Artificial Intelligence

The Accuracy of ChatGPT in Answering FAQs, Making Clinical Recommendations, and Categorizing Patient Symptoms: A Literature Review

John Geracitano, MS[1], Brittney Anderson, MS[2], Melissa Coffel, MS[3], Myles Rosenzweig, BA[4], Spencer D. Dorn, MD[5], Saif Khairat, PhD, MPH[1,3], Jamie Conklin[6]

[1] Carolina Health Informatics Program, University of North Carolina at Chapel Hill, [2] School of Public Health, East Carolina University, [3] School of Nursing, University of North Carolina at Chapel Hill, [4] Gillings School of Global Health, University of North Carolina at Chapel Hill, [5] School of Medicine, University of North Carolina at Chapel Hill, [6] Clinical Nursing Librarian in the Health Sciences Library at the University of North Carolina at Chapel Hill

Keywords: ChatGPT, clinical decision-making, clinical recommendations, FAQs, patient questions, patient symptoms

https://doi.org/10.63116/VXUL2925

Advances in Health Information Science and Practice

Vol. 1, Issue 1, 2025

Background

ChatGPT is a popular open-source large language model (LLM) that uses supervised learning to create human-like queries. In recent years, ChatGPT has generated excitement in the medical field. However, its accuracy must be carefully evaluated to determine its usefulness in patient care. In this literature review, the authors examine whether ChatGPT can accurately answer frequently asked questions (FAQs) from patients, make clinical recommendations, and effectively categorize patient symptoms.

Methods

A database search in PubMed was conducted using the search terms “ChatGPT,” “accuracy,” and “clinical decision-making,” yielding 122 unique references. Two screening stages resulted in 9 studies that met the evaluation criteria for this review.

Results

Analysis of 9 studies showed that while ChatGPT can answer FAQs, offer recommendations, and categorize symptoms in less complicated scenarios, its clinical accuracy ranged from 20% to 95%. ChatGPT may be helpful in specific clinical scenarios; however, its variable accuracy makes it unsuitable as a stand-alone point-of-care product.

Conclusions

ChatGPT is only adept at providing generalized recommendations when individual patient care is more suitable. Further research is needed to identify where ChatGPT delivers the most accurate responses and how it can supplement traditional care.

Chat Generative Pre-Trained Transformer (ChatGPT) is an artificial intelligence (AI) Chatbot developed by the commercial vendor OpenAI and one of the most popular large language models (LLM).[1,2] ChatGPT is defined as an open-source artificial LLM that can be trained using supervised learning on various topics. It has problem-solving capabilities and can respond to queries with human-like answers.[1,3] ChatGPT was the first LLM to achieve the passing threshold of 60% on the United States Medical Licensing Examination (USMLE) without specialized input from humans, demonstrating an ability for comprehensive reasoning that has contributed to increased trust and confidence in integrating LLMs in medical settings.[4] In the USMLE-specific study,[4] Kung et al input 350 publicly available and validated test questions into ChatGPT version 3.0, which was not trained on the test question dataset, showing a potential to incorporate LLMs into medical education about clinical decision-making.

Laypeople and healthcare professionals increasingly use ChatGPT to answer medical questions, triage patient symptoms, and provide treatment recommendations, potentially improving patient outcomes.[5,6] Additionally, studies have shown that ChatGPT responds to patient questions with more empathy and quality than some clinicians.[7] While implementing ChatGPT in healthcare can potentially improve patient outcomes and reduce clinician workloads, its potential functional limitations must be considered, such as errors and misinterpreting complex medical cases.[8]

Frequently asked questions (FAQs) in healthcare address common patient concerns and provide clinical recommendations, aiding in symptom categorization for chronic disease management.[9,10] Research shows that sections for

FAQs enhance patient education, self-management, and adherence to treatment plans.[11-13] Furthermore, digital tools and clinical decision support systems (CDSS) integrated with FAQs help healthcare providers with evidence-based recommendations, improving patient outcomes.[14-16] FAQs also support telemedicine by addressing common queries during remote consultations, enhancing patient-provider communication and satisfaction, and promoting patient-centered care.[17-19]

While many have examined ChatGPT's accuracy in specialty care, its utility and accuracy in clinical decision-making remain largely unknown. This literature review aimed to explore and evaluate ChatGPT's potential in future healthcare settings by assessing its accuracy in answering FAQs, categorizing patient symptoms, and making clinical recommendations.

METHODS

This literature review aimed to identify research articles related to the accuracy of ChatGPT in clinical decision-making. The search terms "ChatGPT," "accuracy," and "clinical decision making" were entered into PubMed's Advanced Search tool. PubMed was selected as the sole database because of its comprehensiveness, and because the scope of our research question did not extend beyond the healthcare field. Studies were included if they tested the accuracy of ChatGPT related to clinical decision-making, were conducted within the United States, were published in the last five years (2019-2024) and were published in English. Systematic reviews, meta-analyses, and opinion pieces were excluded. The database search yielded 125 results. After removing three duplicates, Covidence (Melbourne, Australia) online screening software was used to facilitate title and abstract screening by one author (BA), leaving 50 studies for full-text screening. With one study unable to be retrieved, 49 articles underwent full-text screening, 40 of which were excluded for nonrelevance (n=35), study type (n=2), and study location outside the US (n=3). The same author (BA) then conducted a thematic analysis of the selected nine articles using a matrix format in a Microsoft Word table with the A-15 Point Checklist as a guide.[20] Figure 1.1 illustrates this sequence in a PRISMA 2020 flow diagram.[21]

A thorough review of 9 included articles was conducted to examine if ChatGPT can (1) accurately answer patient FAQs, (2) make clinical recommendations, and (3) effectively categorize patient symptoms. The categorization of patient symptoms includes aspects like the clinician's diagnosis and its relevance in prioritization, such as triaging.

RESULTS

ANSWERING FREQUENTLY ASKED QUESTIONS

One approach to determine the accuracy of ChatGPT in specialty care is evaluating the propriety of its responses to patients' FAQs regarding future care. In this review, specialty care refers to medical specialists such as oncologists instead of family practitioners.

Of the nine studies examined, three[22-24] investigated ChatGPT's ability to respond to FAQs related to surgical procedures, resulting in similar findings despite varying methodologies. Mika et al's[22] focus on preoperative total hip arthroplasty (THA) inquiries found that while generally factual, 70% of ChatGPT's responses needed moderate clarification, while 20% were evaluated as excellent by the research team. Dubin et al[23] compared ChatGPT's THA responses to those of arthroplasty-trained nurses. ChatGPT outperformed the nurses by two percentage points, providing appropriate answers with accuracy levels of 95% to 93%.[23] Furthermore, 79% of patients expressed uncertainty about trusting AI, but 69% preferred ChatGPT responses over those from nurses.[23] Lastly, Li et al[24] found that ChatGPT 4.0 provided comprehensive and human-like answers to perioperative abdominoplasty FAQs but lacked the individualized nature required in plastic surgery. The authors concluded that ChatGPT's answers cannot replace personalized advice offered by licensed plastic surgeons.[24] However, ChatGPT may be helpful in the initial consultation of patients interested in abdominoplasty.[24]

Across these three studies,[22-24] ChatGPT demonstrated the ability to provide mostly accurate and coherent answers. However, the methodologies implemented varied in question selection and evaluation criteria. In the Mika et al study,[22] researchers assessed accuracy using a qualitative rating scale for 10 questions from well-known healthcare institutions. This study's subjective nature is important, as accuracy was not clearly defined, nor were there other metrics considered outside of the qualitative scale.[22] Conversely, Dubin et al[23] had orthopedic surgeons gauge the accuracy of responses between ChatGPT and nurses, and they also included patient perception. Having evaluations performed by true subject matter experts like the surgeons in Dubin et al's study[23] strengthen the significance of the results, as they are trustworthy resources. Similarly, Li et al's[24] approach considered the comprehensiveness and effectiveness of responses through the evaluation by plastic surgeons. Eight commonly asked questions related to abdominoplasty were presented to ChatGPT, and responses were provided regarding accessibility, informativeness, and accuracy. Plastic surgeons crafted the questions following guidelines implemented by the American Society of Plastic Surgeons (ASPS).[6]

The variance in these methodologies is important to note, given their narrow scope of questions and subjective evaluation. An increase in the number of questions posed to ChatGPT could alter the perceived accuracy of responses.

MAKING CLINICAL RECOMMENDATIONS

ChatGPT is expected to make accurate recommendations for varying clinical indications. Accuracy of clinical recommendations is critical for patients seeking care from all types of healthcare providers. Of note, ChatGPT prefaces its answers with a disclaimer indicating that it cannot give medical advice before making clinical recommendations.

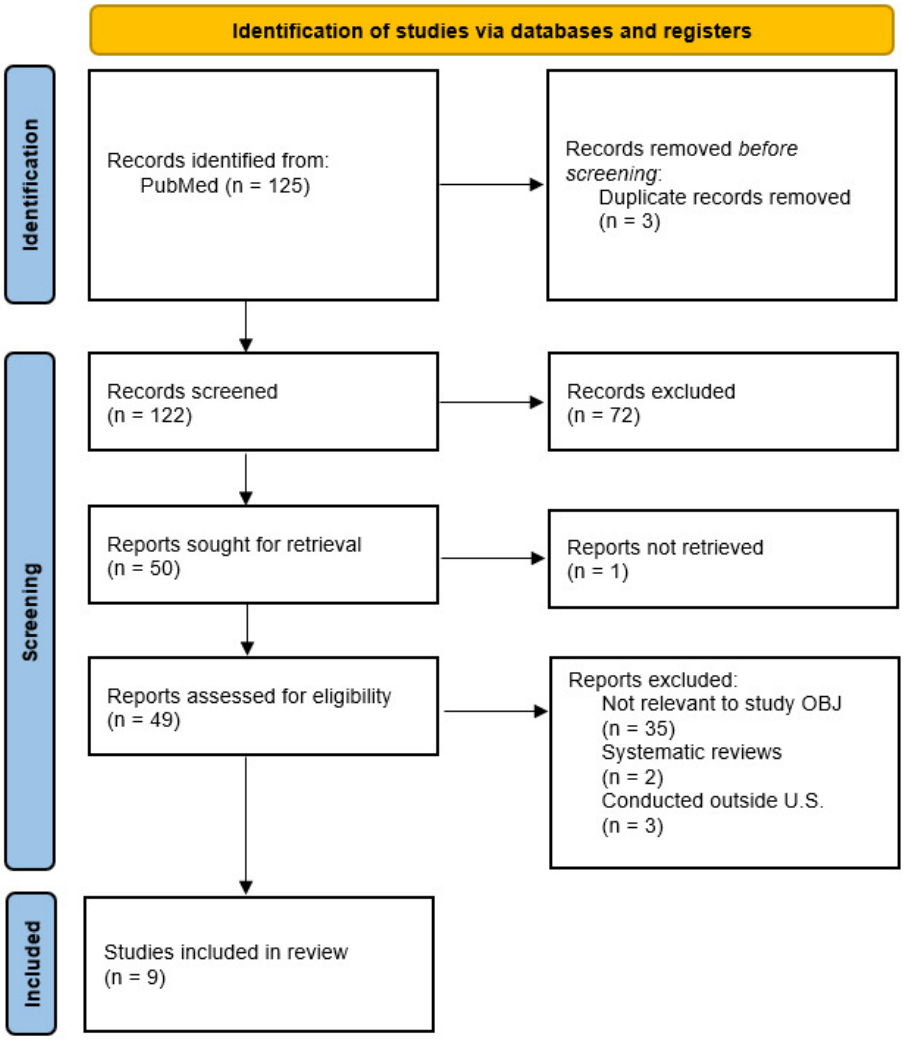

Figure 1.1. PRISMA Diagram for Study Identification

Three studies[25-27] focused on ChatGPT's ability to provide clinical recommendations. Xie et al[25] evaluated ChatGPT's potential to act as a clinical assistant for preoperative rhinoplasty patients, finding that while it effectively outlined the risks, benefits, and outcomes, it had difficulty distinguishing the technical nuances between open and closed rhinoplasty procedures, illustrating an inability to provide individualized recommendations. Similarly, Hermann et al[26] found that ChatGPT 3.5 excelled at cervical cancer prevention and survivorship recommendations, with 91.7% and 93.8% correctness ratings respectively, yet struggled with correct responses in treatment (79.4%) and diagnosis (33.3%). Gajjar et al[27] performed a similar evaluation that compared answers from versions 3.0, 3.5. and 4.0 of ChatGPT within the context of neurosurgical recommendations (spine, cranial, and general neurosurgery), with findings showing similar levels of accuracy across all versions. Of note, version 3.5 was marginally higher in ratings accuracy and helpfulness, while 4.0 had the highest understandability. However, all versions were rated as difficult to read using the Flesch Reading Ease test, which is a factor for users with low health literacy.[27] These studies show ChatGPT's ability to provide generally accurate clinical recommendations, although its effectiveness varies by medical specialty.

A broad range of methodologies and metrics were implemented in the three studies that evaluated clinical recommendation accuracy.[25-27] Xie et al[25] presented nine questions to ChatGPT from hypothetical preoperative rhinoplasty patients to determine whether it could select and explain the most appropriate surgical procedure. Plastic surgeons evaluated its responses. Hermann et al[26] posed 64 questions to ChatGPT related to cervical cancer prevention, diagnosis, and treatment, which gynecologic oncologists assessed for their comprehensiveness and accuracy. The wide range of questions sheds light on ChatGPT's accuracy across clinical categories, as indicated by its high ratings for prevention and survivorship but lower accuracy in treatment and diagnosis. Finally, Gajjar et al[27] incorporated the following assessment metrics across the three ChatGPT versions: accuracy as rated by neurosurgeons and nurses, actionability from the Patient Education Materials Assessment Tool, and readability using the Flesch Reading Ease test. That study's broad inclusion metrics revealed the challenges of understandability from a patient perspective,

which include factors such as health literacy and accessibility in general.

CATEGORIZING PATIENT SYMPTOMS

Three studies[1,3,28] analyzed ChatGPT's ability to categorize or triage patient symptoms. Rao et al[1] assessed the radiologic decision-making of ChatGPT in classifying breast cancer and breast pain clinical indications and recommending appropriate imaging modalities. That study's use of ChatGPT versions 3.5 and 4.0 to interpret clinical scenarios showed a marked improvement in accuracy between versions.[1] ChatGPT 4.0 had an accuracy of 98.4% compared with version 3.5's 88.9% for breast cancer screening recommendations.[1] Similarly, for breast cancer pain recommendations, version 4.0 again surpassed version 3.5 with 77.7% and 58.3% accuracy ratings, respectively.[1] Given that breast cancer is the leading cause of death for women, the accuracy of categorization of breast images is essential.

Ayoub et al[28] sought to determine ChatGPT's triage and diagnostic abilities across cardiology, pulmonology, and neurology, with clinicians assessing performance across multiple metrics. ChatGPT showed a high level of accuracy in differential diagnosis at 88%, although its responses were often incomplete, highlighting a capability gap in clinical reasoning.[28]

Lastly, Dabbas et al[3] explored ChatGPT's capacity for neuro-localization, presenting 47 case scenarios with definitive answers. Seven neurosurgeons graded ChatGPT's responses using a 5-point scoring system, determining that ChatGPT provided "completely correct" responses 69.6% of the cases and 15.2% as "mostly correct."[3] There were six instances where researchers had to give hints to ChatGPT for it to formulate accurate responses.[3] Despite needing assistance, ChatGPT had a basic understanding of neurology and can be utilized to support clinical decision-making.[3] However, the authors highlighted that ChatGPT should not be a primary source for patient care.[3]

DISCUSSION

Based on prior research, ChatGPT has shown varying levels of performance across different medical specialties. Its accuracy ranged from 20% to 95% across the nine selected studies from this literature review,[1,3,22-28] illustrating an inability to autonomously make clinical decisions, triage patient symptoms, or answer patient questions (FAQs). However, in specific settings, ChatGPT showed that it can be a valuable tool to supplement clinical decision making within traditional care.[3,22-25,27,28]

Studies by Mika et al[22] and Ayoub et al[28] revealed that ChatGPT demonstrated notable accuracy rates, particularly in providing differential diagnoses in cardiology and answering FAQs about THA. However, while ChatGPT received high ratings for diagnostic accuracy, it struggled with the completeness of its responses.[22,28] This variability indicates that ChatGPT can be a helpful diagnostic support tool but is not yet suitable for comprehensive clinical decision-making without human oversight. These performance variations highlight the need for domain-specific optimization to ensure that ChatGPT's knowledge base remains accurate and reliable. This aligns with prior literature that found that while AI can assist in diagnosis, human expertise remains crucial to accurately interpreting complex or ambiguous cases.[29]

Patient trust was a pervasive, sometimes inconsistent, factor in the reviewed studies. Dubin et al's[23] research indicated that a majority of patients were more inclined to follow ChatGPT's postoperative instructions, as it provided more accurate answers regarding THA than arthroplasty-trained nurses.[4] This emphasized that in specific scenarios, patients may be more receptive to recommendations from AI over humans. However, nearly 80% of patients remained unsure about fully trusting AI-based recommendations.[23] This finding underscores the complex relationship of patients placing complete trust in technology, which can be influenced by the presentation of information rather than the source's inherent reliability.[9] This dynamic has significant implications for integrating AI tools in patient care settings, where establishing trust is paramount. Moreover, the Food and Drug Administration (FDA) has not yet authorized the use of LLMs in healthcare, which might improve AI's credibility and help foster trust in AI tools.[29]

The successful integration of AI in healthcare settings depends on recognizing its role as a decision-support tool rather than a replacement of healthcare providers. Studies examining ChatGPT in specialized fields such as neurology[3] and gynecologic oncology[26] have shown that ChatGPT's performance is satisfactory but not sufficient to replace specialists. For example, ChatGPT's accuracy in answering questions related to gynecologic oncology was only about 53% for comprehensive responses, which falls short of the standards required in terminal illness contexts.[3,26] This was further represented in Dabbas et al's study, in which ChatGPT required hints from researchers to achieve higher accuracy.[3] Given that the researchers had to aid ChatGPT in decision-making, it is not yet suitable to provide recommendations on its own. Moreover, another study[28] found that ChatGPT lacked the functionality to provide a complete differential diagnosis and will require further input from clinicians before it can benefit patient care. To make AI tools more reliable, future developments should focus on engaging end-users in AI development and enhancing knowledge in these fields while incorporating real-time feedback from clinical experts to improve learning algorithms.

In addition to enhancing the technical capabilities of AI tools, disparities in digital skills remain a significant weakness in these technologies. Current AI tools often provide similar responses regardless of users' digital literacy levels. Gajjar et al's[27] analysis of ChatGPT's responses in neurosurgery revealed that although ChatGPT could deliver highly accurate information, the complexity of its language often created barriers for patients with lower health literacy. This is an obstacle to equitable healthcare access, as the benefits of AI support tools may not be evenly distributed across all patient populations. There is a need for tai-

lored and accessible information to ensure that these technologies do not further exacerbate healthcare disparities.

This review has found that ChatGPT shows promise for use as a supplementary tool in the healthcare setting. However, further exploration of how ChatGPT can be effectively utilized in non-specialty, generalized healthcare settings and clinical cases could offer valuable insights.[24]

RECOMMENDATIONS FOR FUTURE RESEARCH

Future research should focus on developing robust and standardized measurement methods to evaluate the effectiveness of AI tools like ChatGPT for different users, including patients and healthcare providers. First, it is crucial to establish comprehensive metrics that assess the accuracy of AI-generated responses and their clarity, comprehensibility, and practical applicability across broad patient populations. These metrics should account for varying levels of digital health literacy to ensure that AI solutions are available and accessible.

Further investigations should focus on longitudinal studies to assess the long-term effects of AI integration on patient outcomes and provider satisfaction. This includes examining whether AI enhances patient trust, adherence to medical recommendations, and the overall experience in healthcare settings. Additionally, future research should explore AI deployment's ethical and psychological implications, particularly in high-stakes environments where decisions can be life-altering. Developing and validating comprehensive frameworks to measure these factors will help ensure that advancements in AI within healthcare are safe, effective, and aligned with the needs of all stakeholders.

Lastly, there is a significant need for transparency in AI tools. Researchers should investigate innovative methods to clarify AI decision-making processes, such as adopting explainable algorithms that offer clear and concise justifications for each recommendation or response generated. The lack of trust and suboptimal use of AI can largely be attributed to the opacity of "black-box" AI systems. Black-box systems do not reveal how a model arrived at a decision, which contribute to a difficulty in explainability and interpretation. This lack of transparency can undermine trust and impede the integration of AI into clinical practice.

LIMITATIONS

The studies included in this literature review used ChatGPT 3.0, 3.5, and 4. As seen in Gajjar et al's[27] study, varying versions of ChatGPT have demonstrated different levels of accuracy. Ideally, one version of ChatGPT should be investigated for comparability of the results.

Another limitation to ChatGPT's accuracy lies in the extent of its training, as its performance improves with more extensive and targeted training. Therefore, ChatGPT may perform better in making clinical recommendations over time, and its responses may lack precision and reliability without adequate training. This could have led to rudimentary results in the studies included in this literature review. More research should be conducted to understand how accurate ChatGPT can be in clinical decision-making over time. Furthermore, future literature reviews in this field should include other databases such as Science Direct or the Cumulative Index of Nursing and Allied Health Literature (CINAHL).

CONCLUSIONS

Results from this literature review indicated that integrating LLMs like ChatGPT in healthcare has promise for improving clinical decision-making and patient care. However, we must still view AI as a supplement for patients and healthcare providers, rather than a replacement. We found that while ChatGPT offers promising accuracy based on the results of previous studies, its limitations underscore the need for human oversight in clinical applications. Future research should develop standardized measures to assess the accuracy and clarity of AI-generated responses, ensuring accessibility for all patient populations. Longitudinal studies are needed to explore AI's long-term effects on patient outcomes and provider experiences. Enhancing transparency and involving end-users and clinical experts in development will help cultivate trust and effectively integrate AI technologies into traditional care.

CONFLICTS OF INTEREST

The authors have nothing to disclose.

ACKNOWLEDGMENTS

The authors would like to acknowledge Jamie Conklin, Clinical Nursing Librarian in the Health Sciences Library at the University of North Carolina at Chapel Hill, for assisting in developing the search strategy used in this review.

AUTHORS' CONTRIBUTIONS

The scope of this research was conceptualized by BA, MC, and SK. BA conducted the search and screening of recorded abstracts. BA screened the full texts of all articles. A draft of the literature review was written by CRB under guidance from JC, MC, and SK. JG conducted major revisions of the manuscript. The draft was reviewed and edited by all authors. A final copy was approved by all authors.

FUNDING

This study was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health (NIH) under Award Number RC2TR004380. The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of NIH, nor does mention of department or agency names imply endorsement by the US government.

Submitted: January 21, 2025 EDT. Accepted: March 15, 2025 EDT.

Figure 2. Summary of Findings

Reference	Aim of the Study	Research Methodology	Main Findings	Implications for practice
Mika et al (2023)[22]	To evaluate whether ChatGPT could appropriately answer FAQs about total hip arthroplasty (THA); accessing ChatGPT's educational potential	10 FAQs were posed to ChatGPT; answers were analyzed for accuracy based on a rating system developed by researchers: 1. Excellent response, not requiring clarification 2. Satisfactory, requiring minimal clarification 3. Satisfactory, requiring moderate clarification 4. Unsatisfactory, requiring substantial clarification	70% of the responses given by ChatGPT were rated as "satisfactory, requiring moderate clarification." 20% of the responses were excellent and did not require any clarification.	There is no clear indication of how the responses were measured for accuracy. ChatGPT was not presented with follow-up questions to the original 10, which does not mimic real-life encounters.
Ayoub et al. (2023)[28]	To quantitively determine how successfully ChatGPT could triage, synthesize differential diagnoses, and create treatment plans for nine common clinical scenarios within cardiology, pulmonology, and neurology	Cross-sectional study: nine hypothetical clinical scenarios were presented to ChatGPT; three common scenarios within cardiology, pulmonology, and neurology with overlapping symptoms; ChatGPT scored on appropriateness, accuracy, completeness of differential diagnosis, usefulness of response, and evaluation; five clinicians scored ChatGPT's responses.	ChatGPT scored highest amongst clinicians for its accuracy of differential diagnosis at 88%; the lowest score was the completeness of differential diagnosis at 82%.	ChatGPT is helpful as a tool but cannot replace human expertise.
Dabbas et al (2024)[3]	To assess ChatGPT's capabilities of answering questions related to neurolocalization.	Case scenarios (n=46) were presented to ChatGPT for response. Case scenarios only had one definite answer; seven neurosurgeons graded ChatGPT's responses using a five-point scoring system.	ChatGPT responses scored 69.6% as "completely correct" and 15.2% as "mostly correct."	Ideally, ChatGPT's responses should be closer to 100% correct for accuracy. Given that the researchers had to aid ChatGPT, it seems like it cannot make recommendations on its own. However, ChatGPT had a basic understanding of neurology.
Dubin et al (2024)[23]	To evaluate ChatGPT's responses to frequently asked questions related to total hip arthroplasty by asking orthopedic surgeons to rate its responses and askingpatients to evaluate ChatGPT's postoperative questions.	**Phase 1:** 60 questions regarding total hip arthroplasty based on actual patients' questions between 2010 and 2023 presented to ChatGPT 3.5 and arthroplasty-trained nurses; responses graded as "appropriate, inappropriate, or unreliable" by orthopedic surgeons. **Phase 2:** Patients responded to how they perceived ChatGPT 3.5's responses.	**Phase 1:** ChatGPT answered 57/60 or 95% of the questions appropriately compared with nurses, who responded 93.3% correctly. **Phase 2**: 69.4% of patients were more comfortable following ChatGPT responses than the nurses; 79% of patients were uncertain if they trusted AI.	ChatGPT performed better than human experts in this study. This could be due to the volume of questions or how ChatGPT was trained; it was concluded that patients trusted ChatGPT over the nurses.
Gajjar et al (2024)[27]	To assess the accuracy of ChatGPT (3.0, 3.5, 4.0) in answering questions regarding neurosurgery, and to assess the understandability and actionability of ChatGPT responses.	Sixty (20 spines, 20 cranial, 20 general) FAQs presented to CHATGPT versions 3.0, 3.5, and 4.0; responses graded by practitioners. New prompts were between questions to prevent Machine Learning. Answers were graded by five board-certified neurosurgeons and five neurosurgery nurses on a scale of 1 to 5; responses also evaluated using Patient Education Materials Assessment Tool (PEMAT).	ChatGPT was more focused on accuracy than helpfulness; ChatGPT-3.5 was more accurate and helpful than version 3.0 and 4.0; ChatGPT responses had more actionability than understandability.	Although ChatGPT can provide accurate responses to neurosurgery questions, these responses may be too complex for patients with low health literacy to understand.

Reference	Aim of the Study	Research Methodology	Main Findings	Implications for practice
Hermann et al (2023)[26]	To assess the accuracy of ChatGPT's answers to questions related to gynecologic oncology.	Sixty-four questions were gathered from clinical websites regarding gynecologic oncology; two gynecologic oncologists rated the responses to questions using the following grading scale: 1. correct and comprehensive 2. correct but not comprehensive 3. some correct, some incorrect 4. completely incorrect; Tiebreakers were used when necessary.	ChatGPT gave "correct and comprehensive" answers to 34 out of 64 questions (53.1%); ChatGPT answered questions related to prevention and survivorship at 91.7%.	Since gynecologic oncology treats patients with terminal illnesses, patients' questions must be answered with the highest accuracy. ChatGPT did not achieve this level of accuracy in this study, so it should not be used to provide treatment recommendations for gynecologic oncology patients.
Li et al (2023)[24]	To assess the ability of ChatGPT-4 to answer common questions related to abdominoplasty.	Eight common questions related to abdominoplasty were presented to ChatGPT; responses were assessed by plastic surgeons for accessibility, informativeness, and accuracy; prevented ChatGPT from learning from previous responses by utilizing a new dialog box.	ChatGPT gave comprehensive and detailed responses; responses were human-like and not overly complex.	This study lacked the quantitative aspect of other studies, making it difficult to determine how researchers arrived at their conclusions. However, the researchers provided a visual representation of ChatGPT's responses, which could improve readers' comprehension.
Rao et al (2023)[1]	To test the accuracy of ChatGPT in determining the most appropriate imaging procedure for various clinical presentations of breast cancer/breast pain.	Fifteen prompts were presented to ChatGPT using both open-ended (OE) and select all that apply (SATA) formats. Utilized ACR Appropriateness Criteria: "usually appropriate," "may be appropriate," and "usually not appropriate" to rate the accuracy of ChatGPT's decision-making.	ChatGPT 3.5 had an accuracy of 88.9% and ChatGPT 4.0 had an accuracy of 98.4% for breast cancer screening recommendations. For breast cancer pain recommendations, the study showed that ChatGPT 3.5 had an accuracy of 58.3% and ChatGPT 4.0 had an accuracy of 77.7%.	ChatGPT may not be trustworthy for clinical recommendations because its accuracy depends on how adequately the chatbot was trained.
Xie et al. (2023)[25]	To assess the ability of ChatGPT to provide accurate responses to hypothetical questions that are typical during rhinoplasty consultation (ie,ChatGPT's ability to act as a clinical assistant).	Nine hypothetical questions were prompted to ChatGPT and evaluated for accessibility, informativeness, and accuracy by plastic surgeons.	ChatGPT could provide coherent information regarding the risks, benefits, and outcomes of rhinoplasty.	While ChatGPT effectively outlined the risks, benefits, and outcomes, it had difficulty distinguishing the technical nuances between open and closed rhinoplasty procedures, illustrating an inability to provide individualized recommendations

REFERENCES

1. Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. *J Am Coll Radiol*. 2023;20(10):990-997. doi:10.1016/j.jacr.2023.05.003

2. De Angelis L, Baglivo F, Arzilli G, et al. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Front Public Health*. 2023;11:1166120. doi:10.3389/fpubh.2023.1166120

3. Dabbas WF, Odeibat YM, Alhazaimeh M, et al. Accuracy of chatgpt in neurolocalization. *Cureus*. 2024;16(4):e59143. doi:10.7759/cureus.59143

4. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198

5. Nov O, Singh N, Mann D. Putting chatgpt's medical advice to the (turing) test: survey study. *JMIR Med Educ*. 2023;9:e46939. doi:10.2196/46939

6. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. *Sci Rep*. 2023;13(1):17885. doi:10.1038/s41598-023-45223-y

7. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med*. 2023;183(6):589-596. doi:10.1001/jamainternmed.2023.1838

8. Baumgartner C. The potential impact of ChatGPT in clinical and translational medicine. *Clin Transl Med*. 2023;13(3):e1206. doi:10.1002/ctm2.1206

9. Epstein RM, Mauksch L, Carroll J, Jaén CR. Have you really addressed your patient's concerns? *Fam Pract Manag*. 2008;15(3):35-40.

10. Reader TW, Gillespie A, Roberts J. Patient complaints in healthcare systems: a systematic review and coding taxonomy. *BMJ Qual Saf*. 2014;23(8):678-689. doi:10.1136/bmjqs-2013-002437

11. Paterick TE, Patel N, Tajik AJ, Chandrasekaran K. Improving health outcomes through patient education and partnerships with patients. *Proc (Bayl Univ Med Cent)*. 2017;30(1):112-113. doi:10.1080/08998280.2017.11929552

12. Peek K, Sanson-Fisher R, Mackenzie L, Carey M. Interventions to aid patient adherence to physiotherapist prescribed self-management strategies: a systematic review. *Physiotherapy*. 2016;102(2):127-135. doi:10.1016/j.physio.2015.10.003

13. Robinson JH, Callister LC, Berry JA, Dearing KA. Patient-centered care and adherence: definitions and applications to improve outcomes. *J Am Acad Nurse Pract*. 2008;20(12):600-607. doi:10.1111/j.1745-7599.2008.00360.x

14. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *npj Digital Med*. 2020;3(1):17. doi:10.1038/s41746-020-0221-y

15. Khairat S, Marc D, Crosby W, Al Sanousi A. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR Med Inform*. 2018;6(2):e24. doi:10.2196/medinform.8912

16. Chen Z, Liang N, Zhang H, et al. Harnessing the power of clinical decision support systems: challenges and opportunities. *Open Heart*. 2023;10(2). doi:10.1136/openhrt-2023-002432

17. Khairat S, Chourasia P, Muellers KA, Andreadis K, Lin JJ, Ancker JS. Patient and Provider Recommendations for Improved Telemedicine User Experience in Primary Care: A Multi-Center Qualitative Study. *Telemed Rep*. 2023;4(1):21-29. doi:10.1089/tmr.2023.0002

18. Pogorzelska K, Chlabicz S. Patient Satisfaction with Telemedicine during the COVID-19 Pandemic-A Systematic Review. *Int J Environ Res Public Health*. 2022;19(10). doi:10.3390/ijerph19106113

19. Khairat S, Pillai M, Edson B, Gianforcaro R. Evaluating the Telehealth Experience of Patients With COVID-19 Symptoms: Recommendations on Best Practices. *J Patient Exp*. 2020;7(5):665-672. doi:10.1177/2374373520952975

20. *Advances in Global Services and Retail Management: Volume 2*. Anahei Publishing; 2021. doi:10.5038/9781955833035

21. Page M, McKenzie J, Bossuyt P, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71

22. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing chatgpt responses to common patient questions regarding total hip arthroplasty. *J Bone Joint Surg Am*. 2023;105(19):1519-1526. doi:10.2106/JBJS.23.00209

23. Dubin JA, Bains SS, DeRogatis MJ, et al. Appropriateness of Frequently Asked Patient Questions Following Total Hip Arthroplasty From ChatGPT Compared to Arthroplasty-Trained Nurses. *J Arthroplasty*. 2024;39(9S1):S306-S311. doi:10.1016/j.arth.2024.04.020

24. Li W, Chen J, Chen F, Liang J, Yu H. Exploring the Potential of ChatGPT-4 in Responding to Common Questions About Abdominoplasty: An AI-Based Case Study of a Plastic Surgery Consultation. *Aesthetic Plast Surg*. 2024;48(8):1571-1583. doi:10.1007/s00266-023-03660-0

25. Xie Y, Seth I, Hunter-Smith DJ, Rozen WM, Ross R, Lee M. Aesthetic Surgery Advice and Counseling from Artificial Intelligence: A Rhinoplasty Consultation with ChatGPT. *Aesthetic Plast Surg*. 2023;47(5):1985-1993. doi:10.1007/s00266-023-03338-7

26. Hermann CE, Patel JM, Boyd L, Growdon WB, Aviki E, Stasenko M. Let's chat about cervical cancer: Assessing the accuracy of ChatGPT responses to cervical cancer questions. *Gynecol Oncol*. 2023;179:164-168. doi:10.1016/j.ygyno.2023.11.008

27. Gajjar AA, Kumar RP, Paliwoda ED, et al. Usefulness and accuracy of artificial intelligence chatbot responses to patient questions for neurosurgical procedures. *Neurosurgery*. Published online February 14, 2024. doi:10.1227/neu.0000000000002856

28. Ayoub M, Ballout AA, Zayek RA, Ayoub NF. Mind + machine: chatgpt as a basic clinical decisions support tool. *Cureus*. 2023;15(8):e43690. doi:10.7759/cureus.43690

29. Warraich HJ, Tazbaz T, Califf RM. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. *JAMA*. 2025;333(3):241-247. doi:10.1001/jama.2024.21451